Handel, a *free-hands* gesture recognition system

Leonello Tarabella

computerART project of ISTI / CNR - Pisa, Italy
Research Area of the Italian National Council of Research
leonello.tarabella@isti.cnr.it
http://tarabella.isti.cnr.it

Abstract. I describe here a real-time vision-based gesture recognition system used in interactive computer music performances. The performer moves his hands in a video-camera capture area, the camera sends the signal to a video digitizer card plugged into a computer. By processing the reconstructed images of the performer's hands in movement the computer detects x-y positions, shape (posture) and angle of rotation of both the hands. Data extracted from image analysis every frame is used for controlling real-time interactive computer music performances. Two approaches, one more formal the other really operative, are presented.

1 Introduction

Modern human-computer interfaces are extremely rich, incorporating devices such as keyboards and mouse and a wealth of advanced media types: sound, video, animated graphics etc. In addition, advanced interaction strategies are being considered. A good example of that is gesture interaction [1,2] where actions of a system are controlled by a series of hand positions or postures. The term multi-modal is often associated with such interfaces to emphasize that the combined use of multiple modes of perception (e.g. visual and tactile) is relevant to the user's interface [3].

An interface, in a conventional sense, might comprise a window system and a mouse through which interaction is possible by taking into consideration specific areas of windows in the computer display. The system allows for very few degrees of freedom at the same time because usually maps the two-dimensional cursor location to *do-it* commands.

The problem with this approach is that it analyses human performance in terms of encoded rules thus forcing a specific behavior on the performer. With this setting, the user is likely to fill a sense of technological awareness and he tends to perceive the machine as his primary interacting partner. In this way each act is performed to communicate the intended information with great details implicitly inhibiting the potential of human effectors in enriching the semantic content of the information to be communicated [4].

Artistic performers usually needs many degrees of freedom to control at the same time in order to communicate their emotions for giving expression to music based on

technology. This can now be achieved by including a computer in the loop between the human physical actions and the musical response, while addressing two basic principles: - holding nothing at all: controllers respond to body position and motion without requiring anything to be grasped or to be worn connected with wires; - sensitive space: controllers sense the player to give the person the strong feeling of being *bathed* in sound [5].

Moving along the guidelines highlighted so far, I have started developing human-computer interfaces by using non-intrusive devices and systems based on the remote sensing of postures of the hands and more generally of the human body [6].

1.1 Background

At the beginning of the '90s I started to tackle the fascinating realm of the real-time control of digital sound and, together with other researchers and collaborators of C.N.R. in Pisa, I realized a number of devices and systems based on the infrared (IR) and the real-time analysis of video captured images technologies: TwinTowers, Light Baton, UV-Stick, Imaginary Piano and PAGe system. The *TwinTowers* is an electronic device based on IR technology consisting of 2 groups of four elements arranged as the vertical edge of two parallelepipeds. After having presented this device many times at technological and artistic level [7,8], I recently developed a new version, also named PalmDriver, consisting of up 8 groups of 4 elements, which works as a standalone device properly equipped with a MIDI OUT port.

Image processing technology has been used for realizing the other systems. The same hardware and the same strategy have been used for implementing them all. A CCD camera is connected to a video grabber card and, whatever the system, the digital image to be analyzed consists of the reconstructed image by means of an algorithm which filters (that is, accepts) those pixels whose luminance is greater than a predefined threshold. Although this algorithm would be not applicable to a generality of images, it is precise enough to distinguish the luminance values of those pixels corresponding to the hands from the rest of the scene. Besides, in order to improve the robustness of the method, the performer dresses in black and has at his shoulders a black background.

The *Light Baton* system has an on-board light LED source on the conductor's baton tip powered by a battery placed in the cork handle. Implemented for conducting a computerized orchestra, this system recognizes those movements of the baton made by the conductor during a live performance that conform to international standards [9].

The *UV-Stick* (and the systems reported in the following) works on the basis of images of object or the hands themselves captured by the CCD camera and lit by a source light placed where usually the camera is placed. In particular, in the UV-stick the source light is an Ultra Violet lamp that gives the stick (a Plexiglas tube 50cm long and 3cm diameter) a suggestive visual impact of a *laser sword*. The extreme points of the stick are used for detecting the barycenter x-y position and its angular rotation [10].

In the *Imaginary Piano* a pianist sits as usual on a piano stool and takes into account an imaginary line at the height where the keyboard of a real piano usually lies: when a finger, or a hand, crosses that line downward, the systems reports proper information regarding the *key number* and a specific message is issued accordingly to *where* and *how fast* the line has been crossed [11].

In the *PAGe* (Painting by Aerial Gesture) system, positions and movements of a performer's hands are recognized in an wide vertical plane; the performer acts as a painter who uses his hands for selecting colors and nuances of color and for actually drawing a picture; movements are performed in the air and the resulting picture is projected on a large video-screen. Beside, special gestures trigger preset sounds which makes operative the paradigm of *synaesthesy in art* [12,13].

These gesture recognition systems produce data-streams used for controlling sound and graphics in real-time. To map information to sound, that is to define how to link data coming from gesture recognition systems to algorithms that generate music, it's up to the composer himself and is related to a specific composition [14].

After the experience gained from the realization of the above-described systems, I tried a formal approach for realizing a general-purpose system able to recognize shape, position and movement of the hands. The basic idea of this approach to gesture recognition was derived from a paper by V.Cappellini [15] based on the Fourier Transform and developed for recognizing bolts and tools sliding on a conveyor belt to be selected and picked up by a mechanical arm.

Anyway, since this method (presented at the ICMC97 [16] and here briefly reported) although stable and elegant, results rather *time consuming* and therefore not fully suited for real-time application such as interactive controlled computer multimedia performances. So, I developed a less formal but more practical and really operative system for the purpose. I'll describe it in paragraph n.3.

2 Formal approach

The digitized image coming from the camera is transformed into a binary matrix where 1's represent those points $p(x_i,y_j)$ whose luminance level is greater than a predefined threshold. The *barycenter*, i.e. the center of mass (x_c,y_c), is given by the weighted mean of rows and columns on the binary matrix as follows

$$x_c = \frac{\sum_i (i \times \sum_j c_{i,j})}{\sum_{i,j} c_{i,j}} \qquad y_c = \frac{\sum_j (j \times \sum_i c_{i,j})}{\sum_{i,j} c_{i,j}}$$

where x_c and y_c are the coordinates of the center of mass and $c_{i,j}$ is the (i-th, j-th) component of the binary matrix so that: $\sum_j c_{i,j}$ and $\sum_i c_{i,j}$ are the count of pixels valued 1 in the i-th row and in the j-th column respectively and $\sum_{i,j} c_{i,j}$ is the total number of pixels valued 1 representing the hand.

Next step consists in constructing a *one-period-signal* by the distances from the barycenter of those points along the contour taken on radii at predefined angular

steps. For searching the second point of each segment (first one being always the barycenter) program searches on lines $Y=mX+q$ for the most distant point of value=1, that is *white*; m is the angular coefficient of radius which changes with step $\Delta\delta$ corresponding to the virtual *sampling rate frequency*.

Since in general the posture of the hand generates non-convex figures, the scanning algorithm just described produces signals corresponding to a *palmed* hand (like ducks feet). However, as experimentally verified, this approximation does not affect the analysis results. More critical is the choice of step $\Delta\delta$*:* when too large the algorithm produces aliased signals and when too small great amount of computation is requested for the FFT.

Experimentally good values have been found to be $\Delta\vartheta = \frac{2\pi}{32}$ and $\Delta\vartheta = \frac{2\pi}{64}$.

The following figure show two different typical postures of the hands, their corresponding one-period-signals constructed as described and the resulting harmonic spectrum computed by the FFT algorithm.

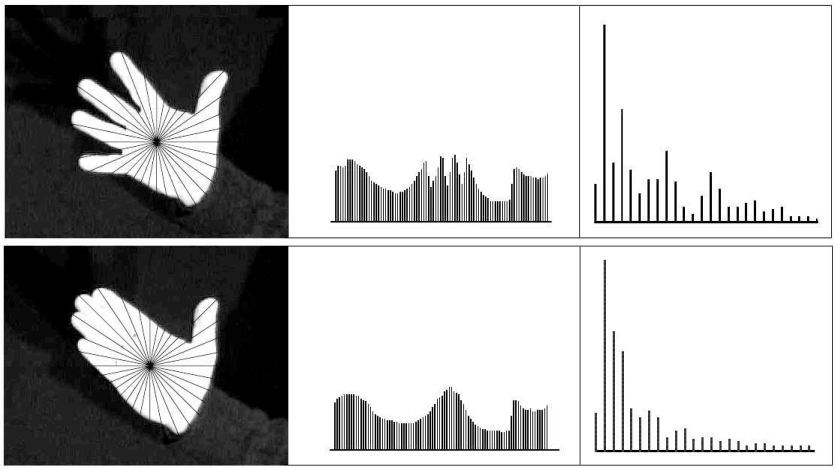

Fig. 1. Postures, one-period-signals and FFT analysis

The harmonic spectrum characterizes very well the posture of the hands and, furthermore, has the very important property of invariance with respect to both rotation and dimension (which changes with the distance from the camera). The result of FFT is input to the actual recognizer that employs an algorithm measuring the distance between n-dimensional vectors.

Let the vector $h=(h_1,....h_n)$ represent the harmonic spectrum derived from the feature associated to a hand's posture and let C be the set of vectors, each representing the harmonic spectrum of a corresponding posture, previously recorded while training the system. The recognizer selects a vector c* from the set C as the harmonic spectrum representative of h such that for all $c=(c_1,....c_n) \in C$ it holds that $\|h-c^*\|_2 \le \|h-c\|_2$ with $\|\ \|_2$ denoting the L2-norm. Rotation comes from the

phase spectrum: in this case only the first component is meaningful since the higher components are simply multiples of the first one.

3 Operative approach

As I said I developed a new system that gets information from postures and positions of the hands. This is less formal but, on the other *hand*, more flexible, faster and more usable thanks to an high number of parameters put at disposal at the same time. Once again, hardware and methodology are based on a video camera which captures the images of the performer's hands, the performer dressing in black on a black background.

Now, instead of recognizing the shape of the hands, it's a matter of taking into consideration the rectangle that *frames* the white spots of the hands. In this way parametric values provided by the real time analysis on the reconstructed dot images deal with dimensions and x,y-coordinates of the center.

The video capture area is converted into a matrix of pixels that is then scanned and analyzed. In the same manner as it happens in the well-known BigEye [17] application, it's possible to define sub-zones where to apply the analysis process. This has two main advantages: the process is faster and at the same time it solves the problem of the presence of the performer's face. Without this facility, a complex and not fully reliable algorithm for filtering out the performer's face should be implemented. The sub-zones where to run the analysis can be dynamically defined.

The whole system is based on ordinary devices such as an analog CCD video camera, a Capsure frame grabber PCMCIA card by IREZ able to convert images with 320x240 pixels at a rate of up to 30 frame/sec and a Macintosh PowerBook G3-500Mhz.

In the following I'll use this terminology: *pane,* i.e. the *defined* sub-area that can be placed everywhere in the capture video camera area with whatever dimensions; *frame,* i.e. the *detected* rectangle that delimits the shape of one hand considered as the reconstructed white spot in memory, therefore defined as *spot-hand*.

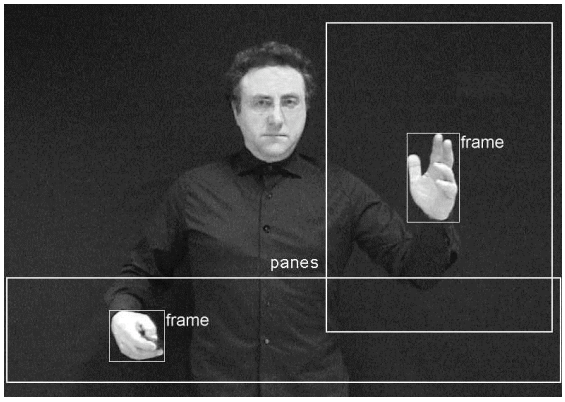

Fig. 2. Typical operative situation

The algorithm that scans and analyses the postures and movements of the hands is simple in principle but, at the same time, it allows a great variety of dynamic figurations truly important for the overall impact on the audience during the performance. In fact it's possible to invent many and new postures and movements to be used in

different musical compositions with any sort of free linkage with the theme and the poetics of the music.

I mean that the great variety of shapes, postures and movements of the hands that can be invented by the composer/performer creativity, can be mapped into the *frame* classes so far described.

3.1 Implementation

Handel has been implemented on a laptop PowerBook Macintosh G3 running at 500Mhz so that the system consisting of CCDcamera +CapsureCard +PowerBook can be considered as a generator of information under the control of the performer's hands gesture. That is, Handel can be considered as a general purpose controller which issues messages of the same type of Midi messages issued by, for example, the old MIDI mixer KAWAI MM-16 used for feeding real time musical commercial products such as MAX.

Actually, I use the data streaming for compositions written in pCM (pureCMusic) I realized and presented in the last years in many conferences and meetings [18,19]. The pCM programming framework gives the possibility to write a piece of music in terms of an algorithmic-composition-based program and of synthesis algorithms also controlled by data streaming from external controllers. Everything is written following the C language syntax, compiled into machine code that runs at CPU speed.

The framework consists of a number of functions for sound processing, for generating complex events and for managing external data coming from standard Midi controllers and/or other special gesture interfaces. For Handel I chose to use the UDP protocol because it has two main advantages: it is faster and makes use of a single small flat cable which plugs directly into the laptops outlets so avoiding the presence of two Midi interface-boxes.

3.2 Analysis

Once a frame is grabbed and converted into a matrix of pixels and stored in memory, the core of the callback routine which implements the functionalities of Handel is invoked; this routine scans the *panes*, search for the spot-hands and, if any, computes and reports dimensions and positions of the *frames*. The *panes* can be dynamically defined by the pCM program/composition as required for different planned musical situations and transmitted to Handel via UDP protocol.

More precisely the callback routine executes the following tasks for each active pane: states the presence/absence of the spot-hand inside pane and, if present, computes the related frame dimensions by scanning the whole pane and storing the higher, the lower, the leftmost and the rightmost pixels coordinates belonging to the spot-hand.

It's also possible, for each pane, to state the step to be used during the scanning: a step value equal 1 means that every pixel is tested; step 2 means that 1/4 out of the totality of the pixels in the pane are tested; step 3 lowers to 1/8 and so on. The step so defines the grid.

As a consequence the algorithm is faster but, on the other side, does not guarantee all the boundary pixels of the spot-hand are tested: the extreme pixels that delimit the frame dimensions can lay on those rows and columns not tested. In this case a wrong value is issued. This error, however, cannot be greater than the value of the step itself and, considering the great advantage gained in terms of velocity, it results quite acceptable. And in any case the user can freely state that.

Since, actually, very often a gesture controlled performance works on thresholds (for example something has to happen when the mass of a spot-hand becomes greater than a predefined value) to lose precision it's tolerable especially when it's a matter of gaining something crucial such as a true real time control.

The formulas and the operative code program, which put them at work, are the well-known formulas for computing the center of mass previously seen. The frame dimensions are simply given by the difference of the coordinates between the extreme points of the spot-hands.

Usually the hands assume posture that show the palm or the back in respect to the CCDcamera and the audience such as those reported in Fig.2. Fingers can be kept closed together or kept in the fist position. Furthermore many combinations of finger-closed/finger-open (such as when counting) can be taken into consideration. With this class of postures the resulting frames are nearly squares so that values to consider are those related to the mass and its position within the pane.

3.3 Angle of rotation

A second class of posture produce *flat* frames, i.e. where one dimension is considerably lower in respect to the other. This is the case where the forearm is placed horizontally and the open fingers point to the camera (mime an airplane flight with thumb and little-finger as the wings)

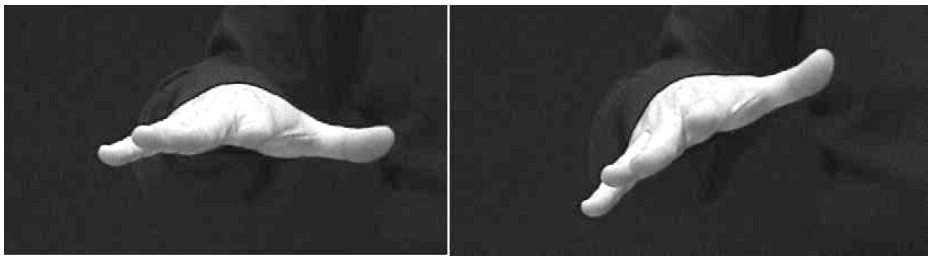

Fig. 3. Flat posture A

or in the posture used in the military salute.

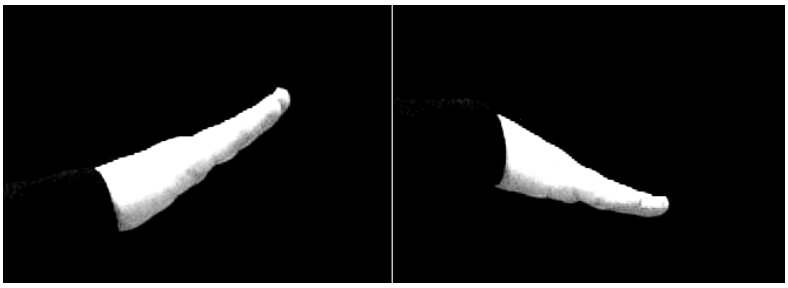

Fig. 4. Flat posture B

With this class of postures the resulting frames are flat and then it makes sense to recognize the angle of rotation. This is computed using the well-known regression-line formulas where x_i and y_i are the coordinates of the white points of the reconstructed spot-hand:

$$slope = m = \frac{s_{xy}}{s_{xx}} \quad \text{with } s_{xy} = \Sigma\ x_i y_i - \left(\frac{\Sigma x_i \Sigma y_i}{N}\right) \quad \text{and} \quad s_{xx} = \Sigma\ (x_i)^2 - \frac{\left(\Sigma\ y_i\right)^2}{N}$$

As a final remark, I want to recall that it's not a matter of recognizing the shape of the hands as seen in the first approach but, rather, that of freely controlling size, position and rotation of the spot-hands which, in turn, change dimensions and rotations of the frames. At the end the hands really control parametric values for giving expression to real time synthesized music.

As summary, these are the information detected by the program.

- spot-hand **presence** in the pane (true/false).
- spot-hand **barycenter** (*x,y)* coordinates
- spot-hand **frame** dimensions (*base,height)*
- spot-hand **angle** of rotation (if meaningful)

and sent via UDP protocol to the computer that runs the pCM program/composition. The number of active rectangles defined by the pCM program/composition can be greater than 2 even if, obviously, those really fully controllable at the same time are only two.

3.4 Future plans

Hardware at the moment in use (PowerBook+Capsure) works very fine. However it is based on obsolete technology: IREZ has not developed the proper drivers for System OS-X so that, should my old PowerBook or the Capsure card go out of order, Handel will be no more usable.

I tried to use webcams based on USB and Firewire protocols but, unfortunately, despite simpler to use, they works with a latency unacceptable in real time applications such as a gesture controlled computer music performance.

For that I'm planning to use the PC-401 standard SBC serie that makes it possible to assembly a very compact special purpose hardware able to: -grab images from

analog cameras; -analyse the spot-hands and produce the related values and -guarantee the requested data rate transmission via USB, Firewire and UDP protocols.

4 Conclusions

Apart the way the whole mechanism works and apart the effective usability of this approach to gesture recognition, I found myself to face the problem concerning the performance visual aspect of *gesturing in the air* in front of the audience. Traditional musical instruments force musicians to assume precise postures of the body and specific movements of the hands in relationship with their mechanic and physical acoustic characteristics.

As a novelty, Handel proposes something where *who* controls and *what* is controlled overlap: the hands are at the same time the instrument and the player. Here, no real instrument exists that forces the performer to specific posture and gesture. Therefore, to be completely free induces to search for a new coherence and elegance to take into account while performing.

I found a first answer to this problem by observing gestures of magicians. In fact, very often -if not always- after my concerts played using the TwinTowers and the ImaginaryPiano, people from the audience freely report to me their impression of having watched a magician beside a musician (the italian word for magician is *prestigiatore* actually a contraction of *presto-digitatore* which means "he who moves fingers quickly"). However, I was not quite satisfied with it because in my performances there is no trick or cheating.

Where I found a deep and valid answer is in *Tai Chi Chuan* that has a considerable variety of movements and posture and, as an Oriental Art, helped me to gain awareness about *unity* (yoga) between body and mind. I want to highlight that I'm not going to try an artistic and/or poetic linkage between the two disciplines just because Tai Chi is not *show* but individual and spiritual research. In any case I feel myself legitimated to use what I learned from Tai Chi about control and coordination of my hands.

5 Aknowledgements

I want to thank Massimo Magrini who, as a very skilled and talented musician and expert in acoustics, electronics and informatics, helped me to choose the necessary hardware and to implement the low level crucial routines for grabbing images into memory and for data transmission via UDP protocol. Also thanks to Ubaldo Swami, my Tai-Chi teacher.

References

1. Bordegoni M., Faconti G.P., Architectural Models of Gesture System, in P.A.Harling and A. Edwards, editors, Proceedings of Gesture Workshop '96, Springer Verlag, pp.61-74, 1996.

2. Rubine D., Specifying gesture by example, in Computer Graphics, V.25(4), pp.329-337. ACM Press, 1991.
3. Hartung K., Münch S., Schomaker L., Guiard-Marigny T., Le Goff B., MacLaverty R., Nijtmans J., Camurri A., Defée I., Benoît C., Development of a System Architecture for the Acquisition, Integration, and Representation of Multimodal Information, A Report of the ESPRIT Project 8579 MIAMI, -WP 3-, March 1996.
4. Duke D.J., Reasoning about gestural interaction, in Proceedings of Eurographics '95, NCC Blackwellm, pp 55-56, 1995,
5. Nagashima Y., Interactive multimedia performance with bio-sensing and bio-feedback, Shizuoka University of Art and Culture & Science Laboratory, in Int. Conference on Auditory Display, 1794-1 Nuguchi, Hamamatsu, Shizuoka 430-8533, JAPAN, 2002.
6. Wexelblat A., An approach to natural gesture in virtual environment, in ACM ToCHI, ACM Press, pp.179-200, 1996.
7. Paradiso, J. Electronic Music: New Ways to Play: IEEE Spectrum Computer Society Press. pp. 18-30, Dec. 1997.
8. Rowe, R. Machine Musicianship. Cambridge: MIT Press. March 2001, pp. 343-353, 2001.
9. Bertini G., Carosi P., Light Baton: A System for Conducting Computer Music Performance, "Interface" (Journal of New Music Research), Lisse, Netherlands, Vol 22 N° 3, pp. 243-247, 1993.
10.Tarabella, L., Magrini M., Scapellato G., Devices for interactive computer music and computer graphics performances: In Procs of IEEE First Workshop on Multimedia Signal Processing, Princeton, NJ, USA - IEEE cat.n.97TH8256, 1997.
11.New Music for Musical Expression, NIME-02, Dublin, Ireland, May 2002.
12.Cardini M., Tarabella L., Wireless, Premio Marconi 2002 per l'arte tecnologica, Circolo Artistico e Università di Bologna, Corte Isolani 7/a, Bologna, 2002.
13.Cardini M., Segno elettronico contemporaneo, International Conference on “Tecnologie e forme nell'arte e nella scienza”, Università degli studi di Salerno, 2003
14.Tarabella,L., Bertini G., About the Role of Mapping in gesture controlled live computer music, in Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR 2003), (Montpellier, France), Lecture Notes in Computer Science (LNCS 2771), Springer Verlag, pp. 217-224, May 2003.
15.Cappellini V., Elaborazione numerica delle immagini, Editore Boringhieri SpA, Torino, 1985.
16.Tarabella L., Magrini M., Scapellato G., A system for recognizing shape, position and rotation of the hands, in Proceedings of th Internationl Computer Music Conference ’98 pp 288-291, ICMA S.Francisco, 1998.
17.Povall R. Realtime control of audio and video through physical motion: Steim's Bigeye: In Proc. Journées d'Informatique Musicale, 1996.
18.Tarabella L., The pCM framework for realtime sound and music generation , in Proceedings of the XIV Colloquium on Musical Informatics, Firenze, Italy, May 2003.
19.Tarabella L., TheObject-Oriented pureCMusic framework, in Proceedings of the International Conference on Understanding and Creating Music, Seconda Università di Napoli, Facoltà di Matematica, Caserta, 2003.